Structural Genomics of *M. Jannaschii*

T. Zarembinski, L.-W. Hung, J. Mueller-Dieckmann, K.-K. Kim*,
H. Yokota, R. Kim, and S.-H. Kim

Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and
University of California, Berkeley, Berkeley, CA 94720, USA

*present address: Gyeongsang National University, Korea

The DNA sequence of a gene does not always yield the cellular function of the protein coded by the gene, but the three-dimensional structure of the protein can be a sensitive indicator of its function. For example, protein structures are classifiable in terms of a finite set of folds associated with a small list of functions. In the early stages of a pilot study of the bacterium *Methanococcus jannaschii*, researchers using protein crystallography at the ALS have determined the structure of a protein of previously unknown function. The structure suggested a small number of possible functions from which biochemical assays were then used to determine the actual function.

The goal of the Human Genome Project is to determine the DNA sequence of the human genome by the year 2005. One of the important objectives of determining the genomic sequences is to understand the cellular and molecular (biochemical and biophysical) functions of all the gene products (i.e., mostly proteins) encoded in the genomes, but the function of a protein cannot be readily inferred from the DNA sequence of a gene unless that sequence is significantly similar to that of a gene whose function is already known. The current estimate of the percentage of genes with gene products of known function varies from approximately 30% to 60%, depending on the genomes sequenced. Furthermore, an even smaller fraction of the genes have gene products with known molecular functions. In structural genomics, we look for clues to the function of a protein in its three-dimensional structure.

Determining the structures of all the gene products of an organism would be an overwhelming task. Fortunately, the current database of protein structures strongly suggests that most proteins are classifiable in terms of a finite set of folds, the "folding basis set," and that each fold may be represented by a small number of biochemical or biophysical functions. Accordingly, large-scale projects to determine the structures of a few representatives from each fold family can provide a foundation for the functional genomics by identifying molecular functions that can be combined with cellular functions derivable from mutational studies, transcription tracking, translation tracking, and interaction tracking.

To this end, we are using the Macromolecular Crystallography Facility (MCF) at the ALS in a pilot study of the fully sequenced model hyperthermophilic archaebacterium *Methanococcus jannaschii*. We have chosen several gene products from this organism—some with known cellular functions but without known molecular functions, and some without any known functions—and have begun to determine their structures. The long-term goal of this project is to determine the structures of representative gene products in order to establish a folding basis set for the approximately 1800 gene products expressed in the microbe. The principal focus on finding a large number of new folds makes phase determination by multiple-wavelength anomalous diffraction (MAD) analysis a necessity.

Early results have already allowed the roles of two "hypothetical" proteins (proteins for which there is no other protein in the database with a gene having a similar DNA sequence and a known function) to be tentatively identified from their structure alone. With data gathered at the MCF, for example, we have determined the structure of hypothetical protein MJ0577 from *M. jannaschii*.

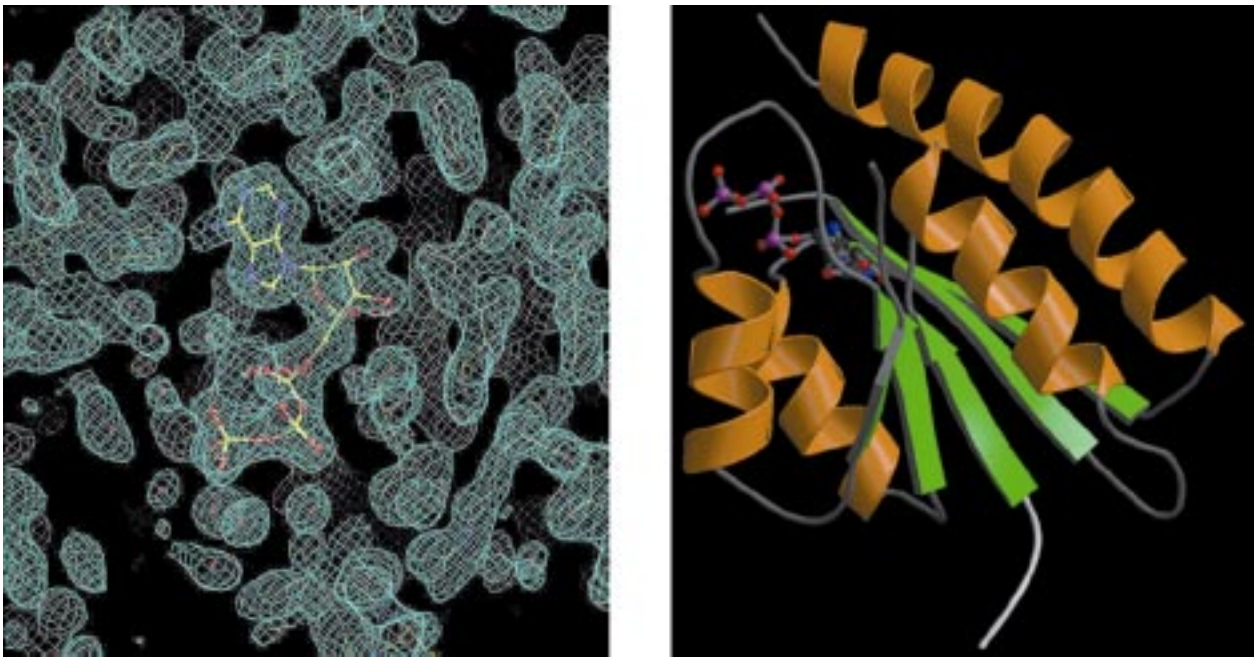

Figure 1. Structure of the hypothetical protein MJ0577 in the hyperthermophilic archaebacterium *Methanococcus jannaschii*, solved at the MCF. (Left) Electron-density map derived from MAD experimental phases clearly shows a bound ATP. (Right) The tertiary structure of MJ0577 is a nucleotide binding fold.

The crystal structure was solved and refined within a few days after data collection was completed. The set of high-quality experimental phases from MAD measurements at the MCF has proven to be the key factor for interpreting and modeling the structures of the protein and ligands. For example, MJ0577 was identified as an ATP-binding protein after examination of the electron density map showed bound ATP.

The discovery of the ATP immediately narrows down the possible biochemical function of this protein. Biochemical experiments showed that MJ0577 has no appreciable ATPase activity by itself. However, when *M. jannaschii* cell extract was added to the reaction mixture, 50% of the ATP was hydrolyzed to ADP in 1 hour at 80˚C. This result indicates that MJ0577 requires one or more soluble components specific to *M. jannaschii* to stimulate ATP hydrolysis, suggesting that this is an ATP-mediated molecular switch analogous to Ras, a GTP-mediated molecular switch that requires GAP to hydrolyze GTP. In these studies, we have shown that MAD experiments can lead to very rapid protein structure determination. Furthermore, in the case of MJ0577, the three-dimensional structure of an unknown protein has provided direct information for functional (biochemical) assignment.

Reference

T. Zarembinski et al., "Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics," *Proc. Natl. Acad. Sci.* **15**, 15189–15193 (1998).

This work was supported by the U.S. Department of Energy, Office of Biological and Environmental Research.

Principal investigator: S.-H. Kim, Structural Biology Division, Ernest Orlando Lawrence Berkeley National Laboratory. Email: SHKim@lbl.gov. Telephone: 510-486-4333.